Simultaneous Speech and Gesture Generation in Embodied Conversational Agents

by

Hao Yan

Thesis proposal for the degree of
Master of Science in Media Arts and Sciences at the
Massachusetts Institute of Technology
November 1999

Thesis Advisor
Justine Cassell
Assistant Professor of Media Arts and Sciences
MIT Media Laboratory

Reader
Alex (Sandy) Pentland
Professor of Media Arts and Sciences, Academic Head
MIT Media Laboratory

Reader
Matthew Stone
Assistant Professor of Computer Science
Rutgers University (New Brunswick)

Abstract

Embodied conversational agent systems are computer interfaces represented by lifelike human or animal characters that are capable of performing believable actions and reacting to human users. Such systems may allow humans to communicate with computers naturally and easily. Humans have long years of practicing communication with other humans, and thus need little training to use such systems.

Face-to-face conversation is about the exchange of information. To achieve this goal, people use a number of communication modalities, such as speech, gesture, facial expression, etc. These modalities are integrated into a conversation unconsciously and without much effort. Therefore, using face-to-face conversation as an interface metaphor, an embodied conversational agent should also have the ability to detect both verbal and non-verbal behaviors from the user, as well as the ability to produce non-verbal behaviors such as gestures while talking in a way that makes sense to the user.

In this thesis, I will build a generation framework that generates prosodically appropriate speech as well as corresponding gestures in real time, with the input of the same underlying knowledge source, discourse structures, and pragmatics. This work will be done in the context of research on an embodied conversational agent. The agent, named Rea (Real Estate Agent), plays the role of a real estate salesperson that interacts with users to determine their needs, shows them around virtual properties, and attempts to sell them a house. Three particular sub-problems will be examined: the distribution of communicative load into different modalities (speech and gesture), the intonation generation in speech utterances, and evaluation of the generation framework.

Index

1. INTRODUCTION 1

2. BACKGROUND 1

2.1 Relationship between Speech and Gesture 1
2.2 Previous Systems that Generate Speech and Gesture 2

3. PROPOSED RESEARCH 2

3.1 Overview 2
3.2 Distribution of Communicative Load across Modalities 3
3.3 Intonation Generation 4
3.4 Evaluation 4
3.5 My Contribution 5
3.6 Completed Work 5

4. MAJOR TASKS AND SCHEDULE 6

5. RESOURCES REQUIRED 6

6. DELIVERABLES 6

7. REFERENCES 7

1. Introduction

Face-to-face conversation is about the exchange of information. In order for that exchange to proceed in an orderly and efficient fashion, participants engage in an elaborate social act that involves behaviors beyond mere recital of information-bearing words. In the following sample conversation, Tim and Lee were sitting in a meeting room waiting for a project meeting. They started the conversation about lodging in Boston area:

(1) Lee: …So, where do you live?

(2) Tim: I live in an apartment in Porter Square. It's about five minutes to the Star Market.

(3) Lee: How is your apartment?

(4) Tim: It's a beautiful place. I like it very much.

In the conversation, many communicative behaviors were employed in addition to the actual speech, either for adding more information or for disambiguation. For example, in the second utterance of (2), by making a walking gesture with his fingers, Tim made it clear that his apartment is five minutes on foot from the star market. In (4), Tim makes an expansive gesture with his hands as a metaphor of "beauty". Meanwhile, the variation of intonation in the production of speech indicates the structure of the propositional content. For example, in the first utterances of (2) and (4) there are pitch accents on "apartment" and "beautiful" respectively, indicating the new and most prominent information conveyed by the utterances. This spontaneous performance, which seamlessly integrates a number of modalities, is given unconsciously and without much effort (Cassell et al., 1999).

Today, the metaphor of face-to-face conversation has been more and more applied to the design of human-computer interfaces. Embodied conversational interfaces are computer interfaces represented by way of lifelike human or animal characters that are capable of performing believable actions and reacting to human users. This kind of interfaces may allow humans to communicate with computers naturally and easily, because humans have long years of practicing communication with other humans, and thus they need little training before they use such a system (Cassell & Stone, 1999).

In order for an embodied conversational agent to have a conversation with the user as Tim and Lee did, the agent should be able to recognize the user's verbal and non-verbal behaviors and give back in real time the same kind of response that humans give in a face-to-face conversation. Research shows that speech and gesture are both integral parts of face-to-face conversation. They are produced simultaneously from an underlying semantic representation. In this thesis, I will build a generation framework in an embodied conversational agent, which generates prosodically appropriate speech as well as corresponding gestures in real time, with the input of the same underlying knowledge source, discourse structures, and pragmatics. I am especially interested in the problem of distributing communicative load into different modalities (speech and gesture), the intonation generation in speech utterances, and the evaluation of the generation process itself.

2. Background

2.1 Relationship between Speech and Gesture

Research has shown strong evidence that there is a close relationship between speech and spontaneous gestures in face-to-face conversations. More than three quarters of all clauses in naturally occurred narrative discourse are accompanied by gestures of one kind or another (McNeill, 1992). Especially when speech is ambiguous or in a noisy environment, people tend to produce more gestures and rely more on gestural cues for understanding (Rogers, 1978). Moreover, spontaneous gestures are synchronized with speech. The most effortful part of a

spontaneous gesture tends to co-occur or occur just before the phonologically most prominent syllable of the accompanying speech (Kendon, 1994). At the semantic and pragmatic level, gestures are also closely related with the accompanying speech. The two communicative channels never convey conflicting information, although they don't always carry the same information. Those concepts that are difficult to express in language, such as manner of actions and simultaneity of two events, may be conveyed by gestures (Kendon, 1994).

Several theories have been developed about the conceptual cause of the close relationship between speech and gesture. Butterworth and Hadar (1989) claim that speech is primary and gesture is a late occurring add-on to language, therefore gesture is not integral to communication. Some claim that instead of being used to communicate information, gesture is the encoding of information in the speaker's mind (Freedman, 1972). McNeill (1992) claim that speech and gesture are both integral parts of face-to-face conversation. They both arise simultaneously from an underlying representation that has both linguistic and visual aspects. This theory explains the strong temporal synchronization constraint on the production of both gesture and speech. It is also suitable to be used as the basis of computational realization of the parallel semantic and pragmatic content in speech and gesture.

2.2 Previous Systems that Generate Speech and Gesture

A few attempts have been made to generate speech and/or gestures in interactive systems. They are complements to earlier works that integrate speech and gesture input such as Put-that-there (Bolt, 1980). Rijpkema and Girard (1991) generated handshapes automatically for an animated character based on the object being gripped in the scene. Perlin and Goldberg (1996) employed rhythmic and stochastic noise functions in developing a system that allows the real-time generation of lifelike behaviors for animated actors. Rickel and Johnson (1999) have their pedagogical agent move to objects in the virtual world that it inhabits, and then based on templates, generate a deictic gesture at the beginning of the verbal explannation that the agent provides about that object. Common in these systems are the lack of consideration of discourse structures and pragmatics to specify non-verbal functions and the ability to allocate communicative load into different modalities.

In Animated Conversation (Cassell et al., 1994), an interaction between two autonomous graphical agents was implemented. It was the first system to automatically produce context-appropriate gestures, facial movements, and intonational patterns for animated agents based on deep semantic representations of information. However, instead of specifying handshapes and hand movements according to semantics, gestures are selected from a canned gesture library. Also, the generation process cannot run in real time.

3. Proposed Research

3.1 Overview

The ultimate goal of this thesis research is to provide in an Embodied Conversational Agent a framework that can generate both prosodically appropriate speech and gestures based on a unified representation of knowledge and discourse context information in real time. This goal can be broken down into three interesting problems: distribution of communicative load across modalities, intonation generation, and evaluation of the generation framework.

This research has both theoretical and practical values. From a research point of view, there are many facets in face-to-face communication, such as content delivering, turn taking, requesting/giving feedback, etc. There are also different theories about the cognitive processes in a face-to-face conversation. By applying those theories into the generation process and observing the output behavior of our agent, we will be able to evaluate those theories. Moreover, it is also a good chance to examine some computational linguistic theories. For example, there are different

theories about how to generate a sentence (or sentences) based on multiple communicative goals. Most of them still stay at the theoretical level. If we can realize those theories in our speech and gesture generation process, we can justify their feasibility. From an interaction design point of view, it is better to have system responses generated based on underlying knowledge and context, rather than canned in some sort of templates, because real-time generation provides more flexibility and variety in the response. For example, to refer to the same internal object ROOM1, the system could generate "a room", "the room" or pronoun "it", depending on the salient information about the object in current context.

This work will be done in the context of research on an embodied conversational agent. The agent, named Rea (Real Estate Agent), plays the role of a real estate salesperson that interacts with users to determine their needs, shows them around virtual properties, and attempts to sell them a house. It has a fully articulated 3D graphical body and communicates using both verbal and non-verbal modalities. She is able to describe features of a house using a combination of speech utterances and gestures, and can also respond to users' verbal (via speech recognition) and non-verbal input (via computer vision). The system uses the SPUD (Sentence Planning Using Description) natural language generation engine (Stone & Doran, 1997) to carry out the response generation task. Figure 2 shows a picture of a user interacting with Rea.

Figure 2: User Interacting with Rea

3.2 Distribution of Communicative Load across Modalities

Gestures do not always carry the same meaning as speech. Those gestures that carry information not present in the simultaneous speech are called complementary gestures, while those that convey the same information as speech are called redundant gestures. One of the first decisions to make in the generation process is, what information should speech and gesture convey respectively, i.e. how to distribute the system's communicative load appropriately into speech and gesture modalities.

Little research has been conducted concerning the above problem. A parallel kind of research has been done by Green (Green et al, 1998) who developed systems that automatically generate presentations consisting of coordinated text and information graphics. The media allocation problem (when to use text and when to use graphics) in those system is solved by applying rules about kinds of information that text and graphics are good at expressing. Cassell and Prevost used information structure, which describes the relation between the content of the utterance and the emerging discourse context, to predict when redundant and complementary information are conveyed across speech and gesture (Cassell & Prevost, 1996). Torres also made good efforts towards formalizing the condition of semantic feature selection by using the concept of semantic structure (Torres, 1997). These works are suitable when generating action descriptions and have not been tested in actual speech and gesture generation system. In this thesis work, I will find more general rules and methods of distributing communicative load across modalities that are appropriate to generate room descriptions and could potentially be applied in other domains as well.

The first step of this work will be studying how people actually divide communicative load by collecting and analyzing video conversation data about real estate. Questions to be answered through this process include: how often gestures convey new information, are there patterns of

information distribution, and how those patterns relate to discourse structure and other pragmatics. Based on the result of data analysis, the next step will be developing a SPUD representation scheme of discourse and pragmatics information and a discourse processing algorithm that can analyze the information used for generation.

3.3 Intonation Generation

Intonation is crucial in conveying the intended meaning in the discourse. In a conversational system, inappropriate intonation selections can be seriously misleading and detract from the intelligibility of what is said. Research shows that the intonation contour coincides with the information structure of discourse (Steedman, 1999, Prevost & Steedman 1994). In particular, a low tone followed by a high pith accent is associated with the theme (or topic) of an utterance, while thehigh pitch accent followed by a low tone is associated with the rheme (or comments).

I will follow (Prevost and Steedman, 1994) for intonation generation based on information structure and contrast modeling. The challenge in this thesis work is, how to use the same kind of knowledge of discourse structure and pragmatics that are used to generate gestures to generate intonation. In addition, current information structure theories are not sufficient to be able to deal with some special cases predicting intonation patterns. For example:

Q1: I am looking for a condo in Boston.

A1: I **have** a condo in Boston.

Q2: I am looking for a place in Boston.

A2: I have a **condo**.

Utterances Q1 and Q2 have the same contribution to the discourse. However, they can elicit two different open questions, whether the addressee has any place in Boston and what kind of place the addressee has in Boston. The resulting answers therefore have totally different intonation contour, “have” in A1 and “condo” in A2 get pitch accents respectively. In this thesis, I will also try to find a solution to this kind of intonation generation problem.

3.4 Evaluation

Unlike many other areas, dialogue systems have proved difficult to evaluate because there is no unique correct dialogue. It is even more difficult to separate and evaluate the response generation part of a system. Walker et al. (1997) described the PARADISE framework for evaluating spoken language dialogue agents in which a single evaluation performance function is defined as a weighted linear combination of both task success (core information transfer) and dialogue costs (efficiency and other qualitative measures). In the area of Natural Language Generation, Mellish and Dale (1998) address the evaluation problem by proposing a scenario – generating summaries from monthly meteorological data – and walking through separably evaluable sub-components and introduce possible measurement in each of them: content determination, document structuring, lexicalization, aggregation, referring expression generation, surface realization. A practical way of evaluating a NLG system is eliciting direct human judgements of fluency. Lester and Porter (1997) developed a two-panel methodology, where texts produced by the system are assessed together with texts generated by a first human panel, while the assessment is done by a second panel.

While the PARADISE type of evaluation is possible and will be valuable in the Rea system, given the limited time of this thesis work, I will follow (Lester and Porter, 1997) to evaluate just the effect and appropriateness of generated speech and gestures. Experiments will be designed in which a human panel will be asked to compare human behavior in a face-to-face conversation with the response that Rea gives. Survey forms will be designed with questions asking if people noticed the agent’s gestures, if they actually got the information conveyed by complementary gestures, etc.

3.5 My Contribution

My major contribution in this thesis work will include:

- Build a generation framework in REA that is capable of generating gestures and speech at the same time (partially completed, see section 3.6).
- By implementing existing discourse theories and evaluating the generation results, I would be able to test the viability of those theories.
- Research rules of people distributing communicative load across modalities and turn those rules into representations that can be used in the generation framework.
- Expand the current discourse model in REA such that it can dynamically produce and update pragmatic features needed for generation.
- Use a comparison survey to evaluate the effectiveness of the generation results.

3.6 Completed Work

I implemented the current generation module on top of SPUD in Rea system. The task of this module is to construct a communicative action that achieves given communicative goals. These propositional goals need to convey domain propositions that encode specified kinds of information about a specified object. The communicative action generated must also fit the context specified by the system's Discourse Manager, to the best extent possible. Figure 1 shows the structure of the simultaneous speech and gesture generation process.

The Structure of Context and the Private&Shared Knowledge together specify a hierarchical organization of types of background knowledge (a group of propositions) that the system and the user share, which defines the common ground (Clark & Marshall, 1981). They also describe the relationship between the system's private information and the questions of interest that information can be used to settle. The Syntactic Frames specifies the syntactic structures of possible types of sentences that can be generated. The Lexicon contains lexical items in speech and constraints on movements in gestures, which are equally treated as lexicalized descriptors (Cassell & Stone, 1999). These four components construct the basic background knowledge base upon which SPUD generator can draw for its communicative content.

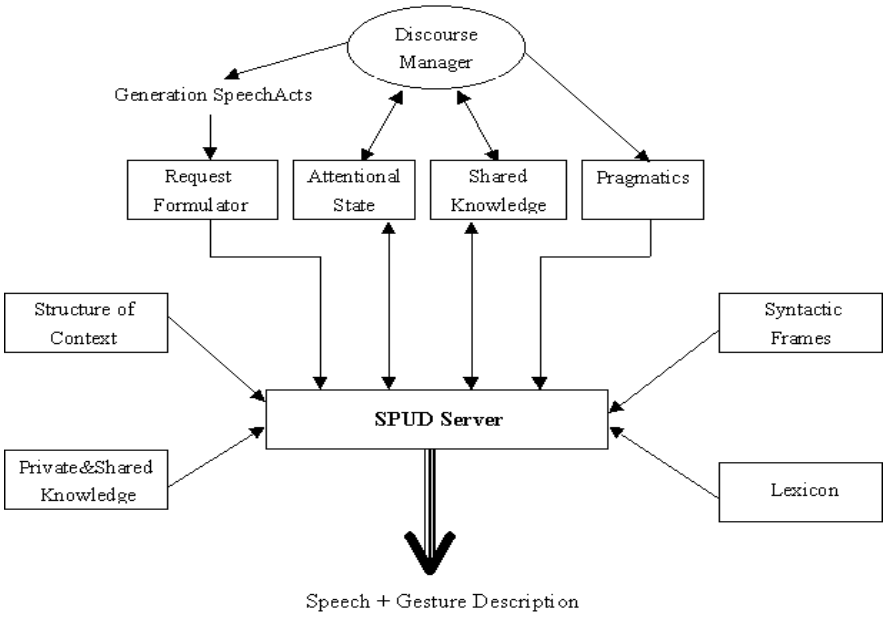

Figure 1: Process of Speech & Gesture Generation

During the conversation, SPUD gets dynamic updates from Rea's Discourse Manager to keep on top of the changing state and context of conversation. These updates include the current attentional state of the discourse (Grosz & Sidner, 1986), shared knowledge update to the common ground, and pragmatics by which SPUD looks to prove before an entry can be used.

An utterance generation process starts when the system's Decision Module sends out a generation command. The generation command is in a speech act format usually in the "Describe(object, aspect)" form. It is first converted by the request formulator into a communicative goal that can be understood by the SPUD generator. SPUD then builds the utterance element-by-element based on the communicative goal, background knowledge base, and the updated context of current conversation. At each stage of construction, SPUD's representation of the current, incomplete utterance specifies its syntax, semantics, interpretation and fit to context (Cassell & Stone, 1999). If a generation process is successful, a speech utterance along with appropriate gesture descriptions are generated.

This implementation can be considered a partial mockup of the generation framework that I am planning to build. There is no intonation generation in it. It also lacks the mechanism of generating pragmatics propositions that will supply SPUD with discourse structure information. The current Discourse Manager in the system does the minimum amount of context tracking and updating. It just keeps a history-list of objects being talked about.

4. Major Tasks and Schedule

- Collecting, coding and analyzing conversation video data in real estate domain. 11~12, 1999.
- Extracting key pragmatic features that determine the generation of gesture and intonation contour. 12, 1999
- Build a discourse model that analyze and updates these features. 01~02, 2000.
- Program a generation grammar in SPUD that uses these features. 02, 2000.
- Conduct a comparison user study to evaluate the effect and appropriateness of the generated speech and gestures. 03, 2000.
- Writing thesis. 03~04, 2000

5. Resources Required

- Human subjects (realtors and customers) for collecting conversation video data.
- A/V devices to analyze videos conversation data.
- A Unix machine that runs SPUD, the natural language generation engine.
- Festival text-to-speech system that can produce intonation variations, provide control tags custom intonation, and have a programming interface.
- CLIPS programming language, which is used to implement the discourse model.
- Resources required for running Rea system, which includes three SGI workstations and two Pentium II Windows-based PCs.

6. Deliverables

- The response generation module in Rea that is capable of generating speech (with appropriate intonation marks) and gestures base on the communicative intent supplied by other modules and the interaction context.
- Discourse data of face-to-face conversation in real estate domain.

- Statistical information about speech, gesture, and their meanings.
- A technical report of specifications about how to use the generation framework.
- A thesis!

7. References

[Bolt, 1980] Bolt, R.A., Put-that-there: voice and gesture at the graphics interface. *Computer Graphics*, 14(3): 262-270, 1980.

[Butterworth and Hadar, 1989] Butterworth, B. and Hadar, U., Gesture, Speech, and Computational Stages: A Reply to McNeill, *Psychological Review*, 96:168-174, 1989.

[Cassell et al., 1994] Cassell, J., Pelachaud, C., Badler, N., Steedman, M., Achorn, B., Becket, T., Douville, B., Prevost, S., & Stone, M. (1994). Animated Conversation: Rule-Based Generation of Facial Expression, Gesture and Spoken Intonation for Multiple Conversational Agents. Proceedings of Siggraph '94, Orlando.

[Cassell and Prevost, 1996] Cassell, J. and S. Prevost. "Distribution of Semantic Features Across Speech and Gesture by Humans and Computers." *Proceedings of the Workshop on the Integration of Gesture in Language and Speech*, 1996.

[Cassell et al., 1999] Cassell, J., Bickmore, T, Campbell, L., Vilhjalmsson, H., and Yan, H., Conversation as a System Framework: Designing Embodied Conversational Agents, to appear in *Embodied Conversational Agents*, Cassell, J. eds, 1999, MIT Press.

[Cassell & Stone, 1999] Cassell, J. and Stone, M. "Living Hand to Mouth: Psychological Theories about Speech and Gesture in Interactive Dialogue Systems." *AAAI 1999 Fall Symposium on Narrative Intelligence*.

[Clark and Marshall, 1981] Clark, H. H., and Marshall, C. R., Definite Reference and Mutual Knowledge. In A. K. Joshi, B. L. Webber, & I. Sag (Editors), *Elements of Discourse Understanding* (pp. 10-63). Cambridge: Cambridge University Press, 1981.

[Freedman, 1972] Freedman, N.,The Analysis of Movement Behavior During the Clinical Interview. In Siegman A. and Pope, B., editors, *Studies in Dyadic Communication*, Pergamon, New York, 1972.

[Green et al., 1998] Green, Nancy and Giuseppe Carenini, Stephan Kerpedjiev, Steven Roth, and Johanna Moore. A Media-Independent Content Language for Integrated Text and Graphics Generation. Proceedings of the Workshop on Content Visualization and Intermedia Representations (CVIR'98) of the 17th International Conference on Computational Linguistics (COLING '98) and the 36th Annual Meeting of the Association for Computational Linguistics (ACL'98). Montreal, Canada, August 15, 1998.

[Grosz and Sidner, 1986] Grosz, B., and Sidner, C., Attention, Intentions, and the Structure of Discourse. *Computational Linguistics*, 12(3), 175-204, 1986.

[Kendon, 1994] Kendon, A., Do Gestures Communicate? A review. *Research on Language and Social Interaction*, 27(3): 175-200, 1994.

[Lester and Porter, 1997] Lester, J. C. and Porter, B. W., Developing and empirically evaluating robust explanation generators: The KNIGHT experiments. *Computational Linguistics* 23. MIT Press, Cambridge, MA, pp. 65-102.

[McNeill, 1992] McNeill, D. (1992). *Hand and Mind: What Gestures Reveal about Thought.* University of Chicago Press, Chicago.

[Mellish and Dale, 1998] Mellish, C. and Dale, R., Evaluation in the context of Natural Language Generation, *Computer Speech & Language*, v12 n4, Oct. 1998, pages 349-373.

[Perlin and Goldberg, 1996] K. Perlin and A. Goldberg. Improv: a system for interactive actors in virtual worlds. In *Proceedings of SIGGRAPH 96*, Computer Graphics Proceedings, Annual Conference Series, pages 205–216, 1996.

[Pierrehumbert and Hirschberg, 1990] Pierrehumbert, J. and Hirschberg, J., The Meaning of Intonational Contours in the Interpretation of Discourse, in Cohen, P. Morgan, and Pollack, editors, *Intentions in Communication*, MIT Press, Cambridge MA, pp. 271-312, 1990.

[Prevost and Steedman, 1994] Prevost, S. and Steedman, M., Specifying Intonation from Context for Speech Synthesis. *Speech Communication*, 15:139-153, 1994.

[Rogers, 1978] Rogers, W.T., The Contribution of Kinesic Illustrators towards the Comprehension of Verbal Behavior Within Utterances. *Human Communication Research*, 5, pages 54-62, 1978.

[Rickel and Johnson, 1999] Rickel J. and Johnson, W. L., Animated agents for procedural training in virtual reality: Perception, cognition and motor control. Applied Artificial Intelligence, 13:343–382, 1999.

[Rijpkema and Girard, 1991] Rijpkema, H. and Girard, M., Computer animation of hands and grasping. *Computer Graphics*, 25(4):339–348, 1991.

[Steedman, 1999] Steedman, M., Information Structure and the Syntax-Phonology Interface, to appear in *Linguistic Inquiry*.

[Stone& Doran, 1997] Stone, M. and Doran, C., Sentence Planning as Description Using Tree-Adjoining Grammar. *Proceedings of ACL 1997*, pages 198--205.

[Torres, 1997] Torres, O.E. (1997). Producing Semantically Appropriate Gestures in Embodied Language Generation. MS thesis, Massachusetts Institute of Technology, Media Laboratory.

[Walker et al, 1997] Walker, M.A. Litman, D., Kamm, C. and Abella, A., PARADISE: A general framework for evaluating spoken dialogue agents. *Proceedings of the 35th Annual Meeting of the Association of Computational Linguistics*, ACL/EACL 97, Madrid, Spain, pp. 271-280